

Execution Models for the Exascale Era

Nicholas J. Wright

Advanced Technology Group, NERSC/LBNL

njwright@lbl.gov

Programming weather, climate, and earth-system models
on heterogeneous multi-core platforms

12-13 September 2012, NCAR

Background and Biases



NERSC Mission: To *accelerate the pace of scientific discovery* by providing high-performance computing, data systems and services to the DOE Office of Science community.

NERSC has over 4500 users in 650 projects that produce about 1500 publications per year!

My Background:

- PhD in Computational Chemistry
- Started in User Services at SDSC – moved to PMaC Lab (Snaveley). Now Advanced Technologies Group at NERSC
- Tools Developer – IPM – www.ipm2.org
- Benchmarking and Performance Analysis, Procurements.

Current NERSC Systems



Large-Scale Computing Systems

Hopper (NERSC-6): Cray XE6

- 6,384 compute nodes, 153,216 cores
- 144 Tflop/s on applications; 1.3 Pflop/s peak

XXXX (NERSC-7): Cray Cascade (Mid 2013)

- >5,000 compute nodes
- >200 Tflop/s on applications; >2 Pflop/s peak



Midrange

140 Tflops total

Carver

- IBM iDataplex cluster
- 9884 cores; 106TF

PDSF (HEP/NP)

- ~1K core cluster

GenePool (JGI)

- ~5K core cluster
- 2.1 PB Isilon File System



NERSC Global Filesystem (NGF)

Uses IBM's GPFS

- 8.5 PB capacity
- 15GB/s of bandwidth



HPSS Archival Storage

- 240 PB capacity
- 5 Tape libraries
- 200 TB disk cache



Analytics & Testbeds



Euclid

(512 GB shared memory)

Dirac 48 Fermi GPU nodes

Magellan Hadoop

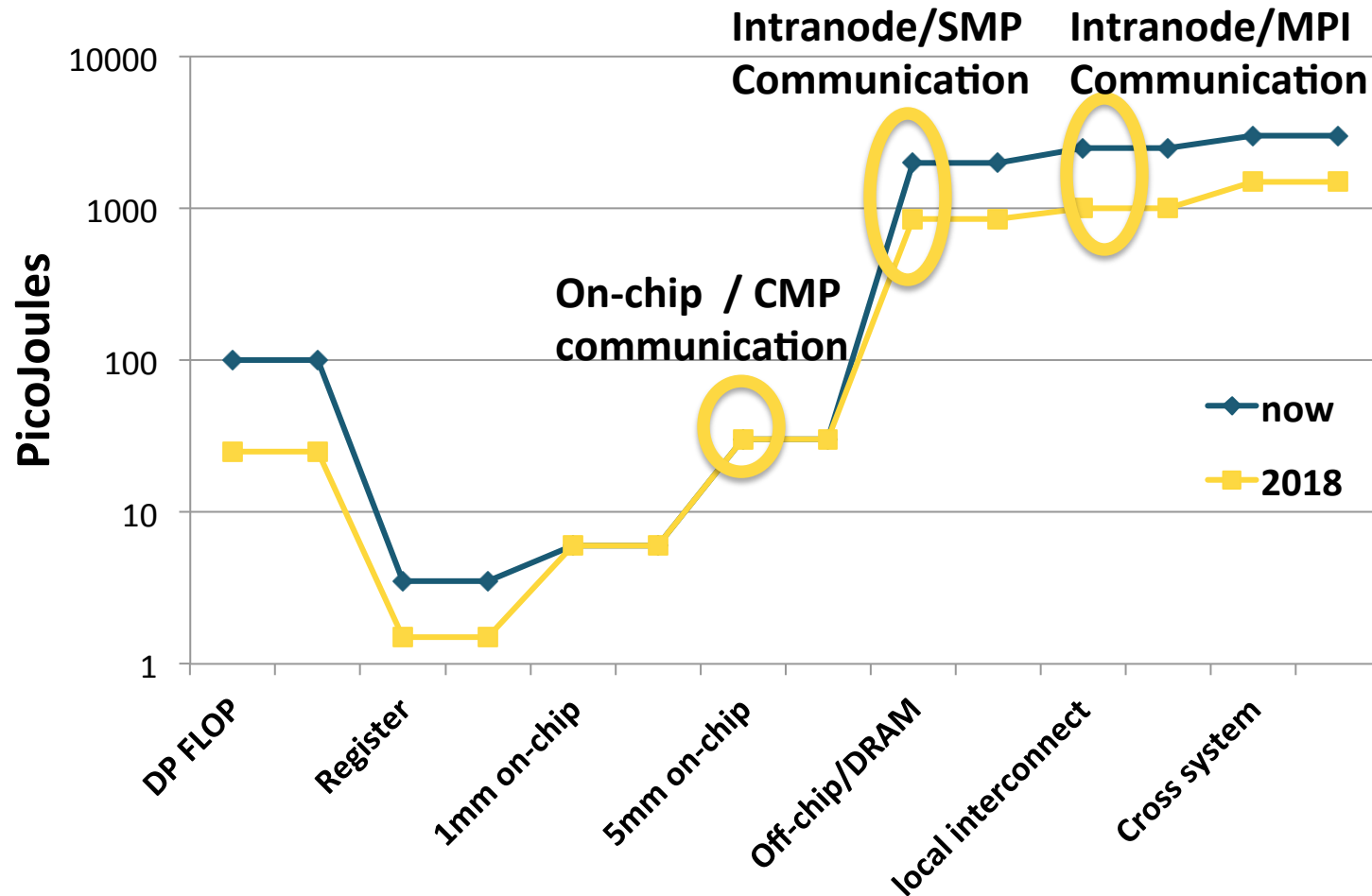


U.S. DEPARTMENT OF
ENERGY

Office of
Science

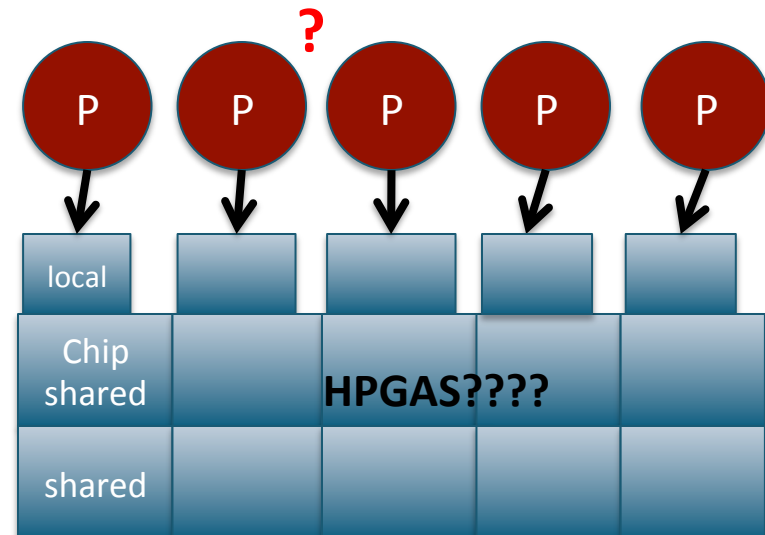
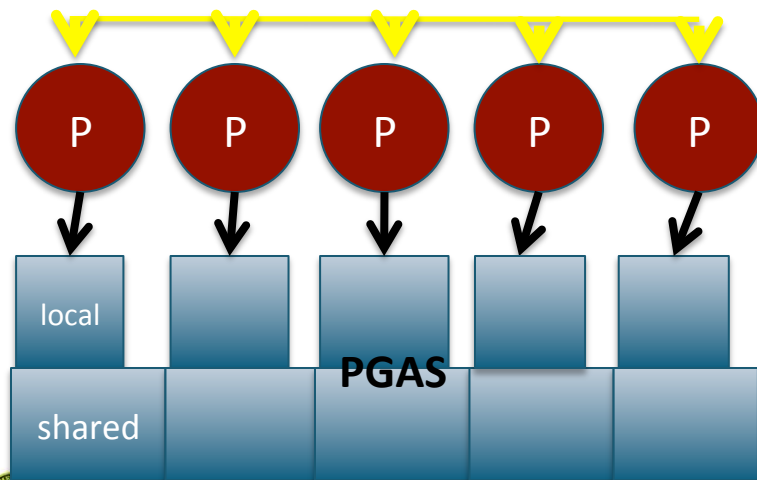
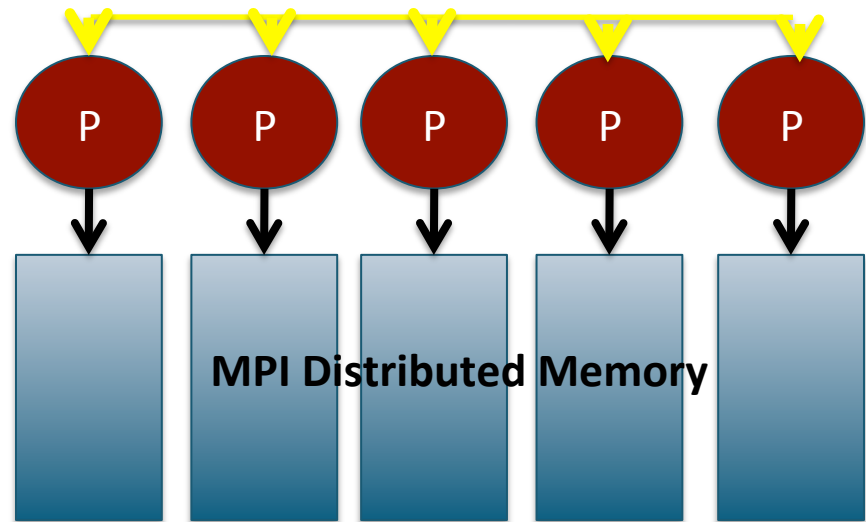
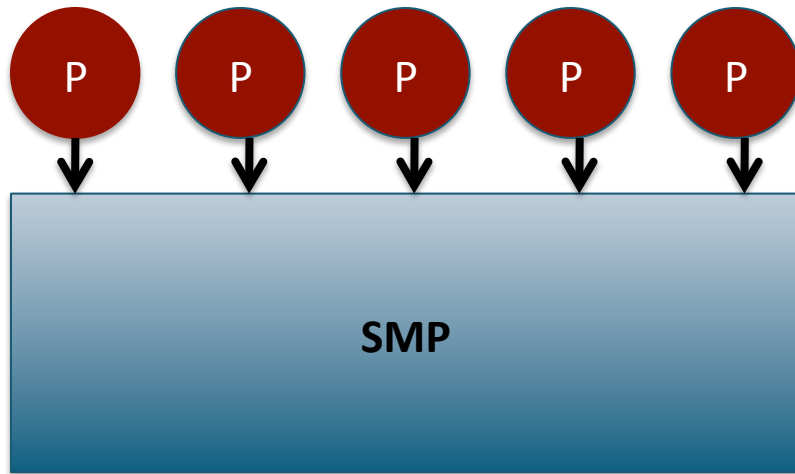


Power: Its all about moving data



Evolution of Abstract Machine Model

(underpinning of programming model)



Challenges to ~~Exascale~~ Performance Growth



- **System power is the primary constraint**
 - Algorithms need to minimize data movement, not flops
- **Concurrency (1000x today)**
 - Need to extract as much parallelism as possible
- **Memory bandwidth and capacity are not keeping pace - likely to see deeper memory hierarchies.**
 - trade off redundant computation for data motion both between nodes and within a node

Challenges to ~~Exascale~~ Performance Growth

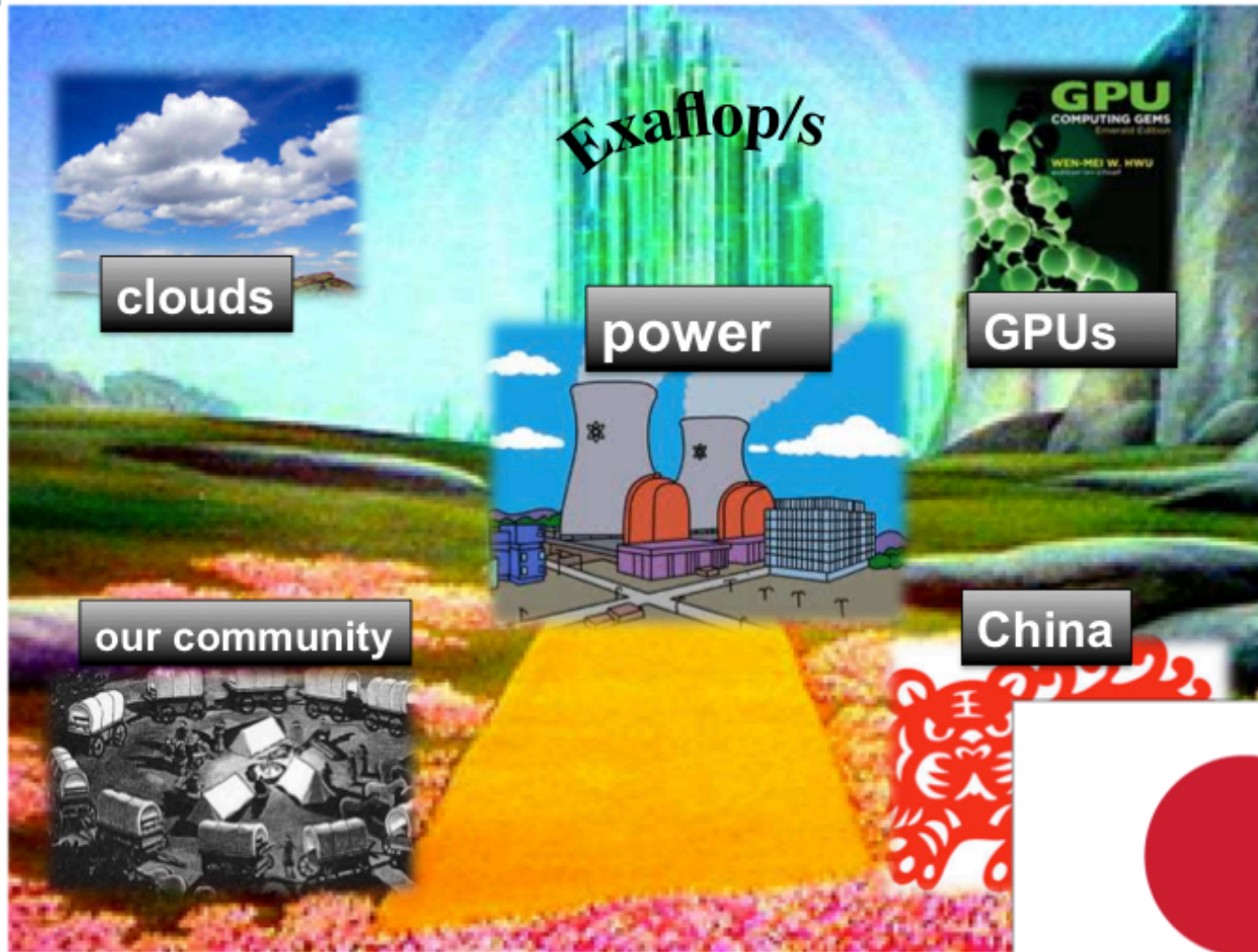


- **Processor architecture is open, but likely heterogeneous. The programming model, heroic compilers will not hide this.**
 - How much will this be a user issue?
- **I/O bandwidth unlikely to keep pace with machine speed**
 - In situ visualization ?
- **Bisection bandwidth limited by cost and energy**
 - Minimize collective communication and synchronization

Horst Simon's Distractions



The Road to Exaflop/s – four distractions and a road block



- **NERSC will install a Cray “Cascade” system in 2013**
 - First all new Cray design since Red Storm; developed for the DARPA HPCS program (including >\$70M from DOE)
 - Intel Processors with > 2PF peak performance
 - New “Aries” interconnect using a “dragonfly” topology
 - 6.5PB storage using Cray Sonexion Lustre appliances
 - Acquisition cost > \$40M
- **Good match for diverse NERSC user needs**
 - Both High-throughput and high-concurrency workloads.
- **Excellent energy efficiency**
 - With benign Bay Area climate, allows chiller-less “free cooling” with PUE < 1.1 for system

Accelerators/GPU's are a passing trend.....



- **Separate memory spaces make programming challenging**
- **More tightly coupled heterogeneous parts are coming - All vendors today have examples of this to different degrees**
 - AMD Fusion
 - Intel Sandybridge
 - NVIDIA project Denver
- **Some people are getting useful science done with GPU's today**
- **Important not to invest time creating non-reusable solutions**
 - Eg Don't rewrite your 1M line code in CUDA.....
- **Wonderful testing framework for exploring issues of extreme parallelism and heterogeneity**

NERSC perspective



"According to NERSC director Kathy Yelick, the lab supports 4,500 users running hundreds of different codes, across many science disciplines and there is concern about forcing all that software to be rewritten for PCIe-based GPUs or Intel MIC devices. "Current accelerators have a separate memory space and are configured as coprocessors rather than first-class cores, both features that we are hoping will change," she explained. "So while we are encouraging users to experiment with low-power processor technology, such as GPUs, in our testbeds, **we do not think the time is right to transition all of the users.**"

http://www.hpcwire.com/hpcwire/2012-07-03/nersc_signs_up_for_multi-petaflop_cascade_supercomputer.html?page=2

Early results from Titan (and NERSC) seem to support this decision



- **Compare 1 XE6 node with 1 XK6 node**
 - GPU \sim 2X price CPU socket
 - GPU 220 Watts CPU Socket 135 Watts
- **Of 11 codes examined only 5 have a speedup of > 1.6**
 - Results will improve with more software tuning and Fermi-> Kepler hardware upgrade
 - Overall not very encouraging
 - Not clear if Intel MIC will obtain significantly better results

Source: http://www.hpcwire.com/hpcwire/2012-07-18/researchers_squeeze_gpu_performance_from_11_big_science_apps.html

Summary



- **Disruptive technology changes are coming**
- **From a NERSC perspective**
 - We want our users to remain productive - Ideally we want them to only re-write their code once
- **Only solve the problems that need to be solved**
 - Don't get distracted by shiny objects
- **Important to understand the cost-benefit tradeoffs and how they apply in *your* situation**
 - For some re-writing in CUDA or OpenACC is worthwhile
 - For some developing processor vendor agnostic techniques maybe important
 - For others waiting for a solution to emerge maybe best

Acknowledgements



- **US Department of Energy Contract No. DE-AC02-05CH11231**
- **NERSC7 team**
 - Tina Declerk, Richard Gerber, Zhengji Zhao, Jason Hick, Lynne Rippe, Jeff Broughton
- **John Shalf, Kathy Yelick**